Understanding LLMs

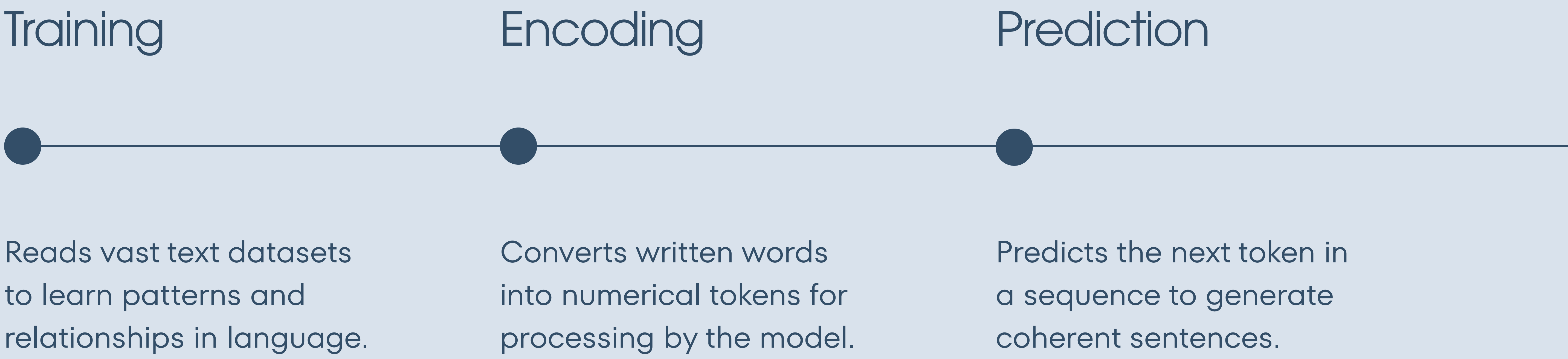
Presented by Waqar Ali

What is an LLM?

Large Language Model *overview:*

A Large Language Model (LLM) is an AI system trained on billions of words from books, websites, and articles.
It learns grammar, facts, and how humans express ideas.
But it doesn’t search the internet — instead, it predicts what comes next based on its training.
That’s why it can generate essays, code, or answers,

How LLMs Work



Training

Reads vast text datasets to learn patterns and relationships in language.

Encoding

Converts written words into numerical tokens for processing by the model.

Prediction

Predicts the next token in a sequence to generate coherent sentences.

What is Prompt Engineering?

Prompt Engineering is the art of talking clearly to the AI.

A prompt is what you type to get your desireable answer from LLM..

Prompt Engineering

Understanding Effective Prompts

*A good prompt combines **role, task, and constraints** to guide LLM responses effectively. Including examples helps clarify expectations and enhances model performance.*

What is Context Engineering

Context Engineering means giving the model the right background.
If you provide facts, examples, and audience details, the model can respond more accurately.
Example: “For college students, explain how ChatGPT learns.”
Keep the context short and focused — if it’s too long, the model may ignore or forget early parts.
Always label sections clearly like SOURCE, TASK, or INSTRUCTION to

Context Engineering

Understanding Context Importance

Context consists of **background information**, examples, and facts. It is vital for effective LLM performance, keeping prompts relevant and concise to avoid exceeding token limits.

Temperature, Top-p, Top-k

Understanding LLM parameters

Temperature

Temperature controls the **creativity level** in responses. A low setting produces focused, factual answers, while a high setting encourages creative and random outputs, affecting overall response quality.

Top-k

Top-k limits the model's choices to the **k most probable tokens**. This stabilizes outputs by ensuring only the highest likelihood options are considered, minimizing the chance of irrelevant responses.

Top-p

Top-p selects tokens based on cumulative **probability thresholds**, ensuring diverse yet relevant outputs. It adapts the selection process dynamically, enhancing the model's flexibility while maintaining coherence in responses.